MIRI

MACHINE INTELLIGENCE
RESEARCH INSTITUTE

ESSENTIAL BACKGROUND

Organization: **Machine Intelligence Research Institute (MIRI)**

Mission: **To ensure that the creation of smarter-than-human intelligence has a positive impact.**

Key Personnel:
Nate Soares (Executive Director)
Eliezer Yudkowsky (Senior Research Fellow)
Benja Fallenstein (Research Fellow)
Patrick LaVictoire (Research Fellow)

Our Research: As more and more human cognitive feats come to be reproduced by artificially intelligent systems, we can expect to encounter a number of unprecedented risks and benefits. For example, our ability to design software is itself a capacity that we may eventually be able to automate; but once AI algorithms are doing the bulk of the intellectual work in designing new AI systems, it is not clear how to establish safety guarantees that apply not only to the programs we write, but also to the programs our programs write.

These and other under-explored questions will only become more pressing as progress in computer science allows AI systems to act with increasing autonomy and efficiency. Although AI systems may not rival humans in social engineering, programming, or scientific ability for many years to come, it seems prudent to begin investigating basic safety issues in advance. There are many organizations focused on improving long-term AI capabilities, but few focused on matching this with long-term progress in security and safety engineering. MIRI exists to help fill that gap.

MIRI's three research objectives are, at present:

- *highly reliable agent design:* how can we design AI systems that reliably pursue the goals they are given?
- *value learning:* how can we design learning systems to learn goals that are aligned with human values?
- *error tolerance:* how can we build AI systems that cooperate with the operators as they improve its design, given that programmer errors are inevitable?

KEY PERSONNEL

Nate Soares EXECUTIVE DIRECTOR

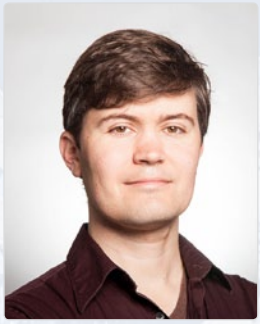
Nate Soares is the director of MIRI's research efforts and day-to-day operations. He joined the MIRI research team in April 2014, quickly earning a strong reputation for his productivity, organization, and strategic insight. Nate is the primary author of most of MIRI's technical agenda, including the overview document "Aligning Superintelligence with Human Interests: a Technical Research Agenda" (2014) and the AAAI conference paper "Corrigibility" (2015). Prior to MIRI, Nate worked as a software engineer at Google.

Eliezer Yudkowsky SENIOR RESEARCH FELLOW

Eliezer Yudkowsky is a decision theorist who is widely cited for his writings on the long-term future of artificial intelligence. His views on the social and philosophical implications of AI have had a major impact on ongoing debates in the field, and his work has heavily shaped MIRI's research agenda. He is the author of the chapter "The Ethics of Artificial Intelligence" (2014, with Nick Bostrom) in *The Cambridge Handbook of Artificial Intelligence*, and of the technical report "Intelligence Explosion Microeconomics" (2013). He is also well known for his writings on human rationality, including "Cognitive Biases Potentially Affecting Judgement of Global Risks" (2008) and *Rationality: From AI to Zombies* (2015).

Benya Fallenstein RESEARCH FELLOW

Benya Fallenstein works on technical problems of self-reference and coordination that arise when formal agents construct or encounter agents similar to themselves, and on gaps in our understanding of logical uncertainty. Benya is also interested in formal verification and in programming languages with integrated proof checkers. Benya is the primary author of "Vingean Reflection: Reliable Reasoning for Self-Improving Agents" and "Reflective Oracles: A Foundation for Classical Game Theory" (both 2015).



Patrick LaVictoire RESEARCH FELLOW

Patrick LaVictoire is a new addition to MIRI's full-time research team. He has made substantial contributions to open problems in decision theory, game theory, and provability theory, and is the lead author of "Program Equilibrium in the Prisoner's Dilemma via Löb's Theorem," a paper presented at AAAI 2014. Patrick's background is in pure mathematics (PhD from UC Berkeley, postdoc at the University of Wisconsin) and applied machine learning (at the app search company Quixey).

Stuart Russell RESEARCH ADVISOR

Stuart Russell is a Professor of Computer Science and Smith-Zadeh Professor in Engineering at the University of California, Berkeley. He is also an Adjunct Professor of Neurological Surgery at the University of California, San Francisco and Vice-Chair of the World Economic Forum's Council on AI and Robotics. He has published over 150 papers on a wide range of topics in artificial intelligence including machine learning, probabilistic reasoning, knowledge representation, planning, real-time decision making, multitarget tracking, computer vision, computational physiology, and global seismic monitoring. With Peter Norvig, he co-authored *Artificial Intelligence: A Modern Approach,* the leading textbook in the field of artificial intelligence.

Nick Bostrom RESEARCH ADVISOR

Nick Bostrom is an Oxford-based philosopher and specialist in emerging technologies. He is the founder and director of the Future of Humanity Institute, an interdisciplinary research center whose researchers regularly collaborate and exchange ideas with MIRI research staff. Nick helped bring long-term risks associated with artificial intelligence to the attention of the public as well as the academic world with his 2014 monograph *Superintelligence: Paths, Dangers, Strategies* and the 2008 anthology *Global Catastrophic Risks*. He has authored over 200 publications, and was listed as one of the top world thinkers by the magazines *Foreign Policy* (ranked seventy-third in 2009) and *Prospect* (ranked fifteenth in 2014).

Max Tegmark GENERAL ADVISOR

Max Tegmark is an MIT cosmologist known for his work on the foundations of physics. He is a Fellow of the American Physical Society, and received Science's "Breakthrough of the Year" prize in 2003 for his work with the Sloan Digital Sky Survey. Of his over 200 publications, nine have been cited more than 500 times. Driven by his interest in investigating fundamental questions in philosophy and science, including the long-term social consequences of new technologies, Max founded the Foundational Questions Institute in 2005. In 2014, he joined Jaan Tallinn and others in co-founding the Future of Life Institute, a Boston-based research and outreach organization that has helped bring leaders in industry and academia together to discuss safety challenges in AI. In 2015, this culminated in a widely supported open letter, "Research Priorities for Robust and Beneficial Artificial Intelligence."

CONTACT SHEET

Address: Machine Intelligence Research Institute
2030 Addison St. #300
Berkeley, CA 94704

Website: www.intelligence.org

Email: contact@intelligence.org

Phone: (510) 859-4381

OVERVIEW OF MEDIA COVERAGE

For more information:

WWW.INTELLIGENCE.ORG